IT & C

ISSN 2821 - 8469, ISSN – L 2821 - 8469, Volumul 2, Numărul 3, Septembrie 2023

Rezumarea automată în inteligența artificială prin învățare nesupravegheată: TextRank

Nicolae Sfetcu

Sfetcu, Nicolae (2023), Rezumarea automată în inteligența artificială prin învățare nesupravegheată: TextRank, *IT & C*, **2**:3, 43-52, DOI: 10.58679/IT78864, https://www.internetmobile.ro/rezumarea-automata-in-inteligenta-artificiala-prin-invatare-nesupravegheata-textrank/

Publicat online: 25.07.2023

Rezumarea automată în inteligența artificială prin învățare nesupravegheată: TextRank

Nicolae Sfetcu[1]
nicolae@sfetcu.com

Automatic summarization in artificial intelligence via unsupervised learning: TextRank

Abstract

Automatic summarization is the process of summarizing a text document with a computer program to create a summary that captures the most important points of the original document. Technologies that can make a coherent abstract take into account variables such as length, writing style, and syntax. Machine learning is a subfield of artificial intelligence dedicated to understanding and building methods that allow machines to "learn". One key phrase extraction algorithm is TextRank, which exploits the structure of the text itself to determine key phrases that appear "central" to the text.

Keywords: summarization, machine learning, unsupervised machine learning, TextRank, LexRank

Rezumat

Rezumarea automată este procesul de sumarizare a unui document text cu un program de calculator pentru a crea un rezumat care să rețină cele mai importante puncte ale documentului original. Tehnologiile care pot face un rezumat coerent iau în considerare variabile precum lungimea, stilul de scriere și sintaxa. Învățarea automată este un subdomeniu al inteligenței artificiale dedicat înțelegerii și construirii de metode care permit mașinilor să ”învețe”. Un algoritm de extragere a frazelor cheie este TextRank, careexploatează structura textului în sine pentru a determina expresiile cheie care apar „centrale” pentru text.

Cuvinte cheie: rezumare, învățarea automată, învățarea automată nesupravegheată, TextRank, LexRank

[1] Cercetător - Academia Română - Comitetul Român de Istoria și Filosofia Științei și Tehnicii (CRIFST), Divizia de Istoria Științei (DIS)

IT & C, Volumul 2, Numărul 3, Septembrie 2023, pp. 43-52
ISSN 2821 - 8469, ISSN – L 2821 – 8469, DOI: 10.58679/IT78864
URL: https://www.internetmobile.ro/rezumarea-automata-in-inteligenta-artificiala-prin-invatare-nesupravegheata-textrank/

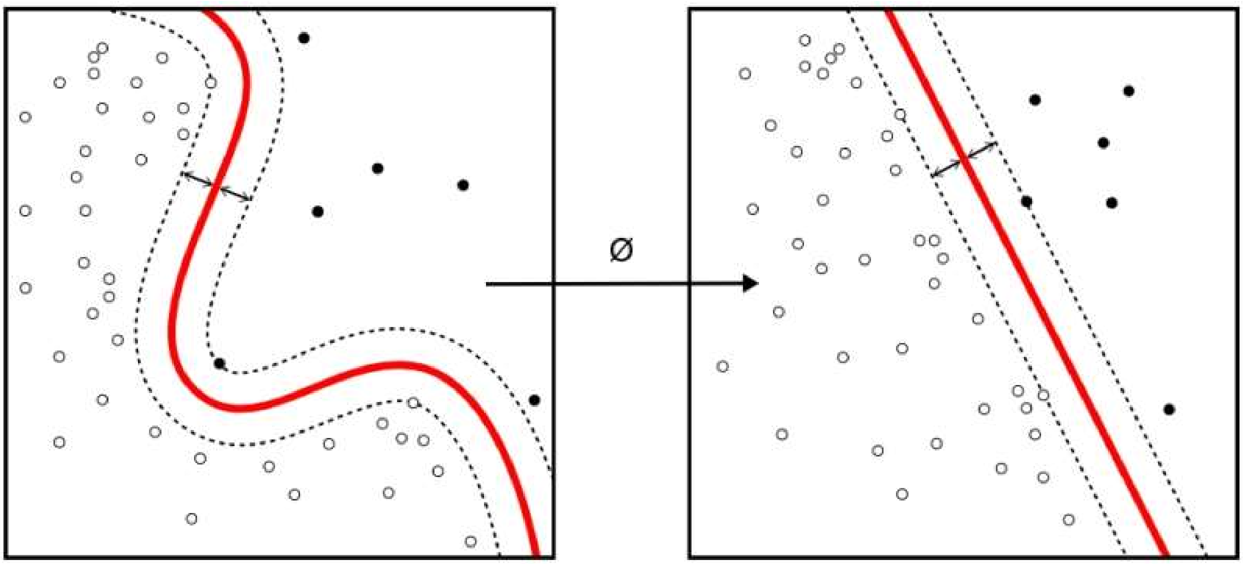
Rezumarea automată

Rezumarea automată (1) este procesul de sumarizare a unui document text cu un program de calculator pentru a crea un rezumat care să rețină cele mai importante puncte ale documentului original. Tehnologiile care pot face un rezumat coerent iau în considerare variabile precum lungimea, stilul de scriere și sintaxa. Rezumarea automată a datelor face parte din învățarea automată și extragerea datelor. Ideea principală a rezumatului este de a găsi un subset reprezentativ de date, care conține *informațiile* întregului set. Tehnologiile de rezumat sunt folosite astăzi într-un număr mare de sectoare din industrie. Un exemplu de utilizare a tehnologiei de rezumare îl reprezintă motoarele de căutare, cum ar fi Google. Alte exemple includ rezumatul documentelor, rezumatul colecției de imagini și rezumatul video. Rezumatul documentului încearcă să creeze automat un *sumar reprezentativ* sau un *rezumat* al întregului document, prin găsirea celor mai *informative* propoziții. În mod similar, în rezumarea imaginilor, sistemul găsește imaginile cele mai reprezentative și importante (sau proeminente). În mod similar, în videoclipurile pentru consumatori a'i dori să eliminați scenele plictisitoare sau repetitive și să extrageți o versiune mult mai scurtă și concisă a videoclipului. Acest lucru este, de asemenea, important, să zicem pentru videoclipurile de supraveghere, în care s-ar putea să doriți să extrageți numai evenimente importante din videoclipul înregistrat, din moment ce cea mai mare parte a videoclipului poate fi neinteresantă, fără a se întâmpla nimic. Pe măsură ce problema supraîncărcării informaționale crește și pe măsură ce cantitatea de date crește, interesul pentru rezumarea automată crește și el.

În general, există două abordări ale rezumării automate: *extracția* și *abstracția*. Metodele extractive funcționează prin selectarea unui subset de cuvinte, fraze sau propoziții existente în textul original pentru a forma rezumatul. În schimb, metodele abstractive construiesc o

reprezentare semantică internă și apoi folosesc tehnici de generare a limbajului natural pentru a crea un rezumat care este mai aproape de ce poate genera un om. Un astfel de rezumat poate conține cuvinte care nu sunt prezente în mod explicit în original. Cercetarea metodelor abstractive este un domeniu de cercetare din ce în ce mai important și activ; cu toate acestea, din cauza constrângerilor de complexitate, cercetarea până în prezent s-a concentrat în primul rând pe metodele extractive. În unele domenii de aplicație, rezumatul extractiv are mai mult sens. Exemple dintre acestea includ rezumatul colecțiilor de imagini și rezumatul videoclipurilor.

Învățarea automată



(Mașinile kernel sunt utilizate pentru a calcula funcții neliniar separabile într-o funcție de dimensiune mai mare separabilă liniar. Credit: Alisneaky/Wikimedia, licența CC0 1.0)

Învățarea automată (2) este un domeniu dedicat înțelegerii și construirii de metode care permit mașinilor să „învețe" - adică metode care valorifică datele pentru a îmbunătăți performanța computerului la un set de sarcini.[1] Este văzut ca un subdomeniu larg al inteligenței artificiale [2].

Algoritmii de învățare automată construiesc un model bazat pe date eșantion, cunoscut sub numele de date de antrenament, pentru a face predicții sau decizii fără a fi programat în mod explicit pentru a face acest lucru.[3] Algoritmii de învățare automată sunt utilizați într-o mare varietate de aplicații, cum ar fi în medicină, filtrarea e-mailului, recunoașterea vorbirii, agricultură și viziunea computerizată, unde este dificil sau imposibil să se dezvolte algoritmi convenționali pentru a îndeplini sarcinile necesare.[4][5]]

Un subset al învățării automate este strâns legat de statisticile computaționale, care se concentrează pe realizarea de predicții folosind computere, dar nu toată învățarea automată este învățare statistică. Studiul optimizării matematice oferă metode, teorii și domenii de aplicare în domeniul învățării automate. Mineritul datelor este un domeniu de studiu conex, concentrându-se pe analiza exploratorie a datelor prin învățare nesupravegheată.[7][8]

Unele implementări ale învățării automate folosesc datele și rețelele neuronale într-un mod care imită funcționarea unui creier biologic.[9][10]

În aplicarea sa în problemele de afaceri, învățarea automată este denumită și analitica predictivă.

Algoritmii de învățare funcționează pe baza faptului că strategiile, algoritmii și inferențele care au funcționat bine în trecut vor continua să funcționeze bine în viitor. Aceste concluzii pot fi uneori evidente, cum ar fi „deoarece soarele a răsărit în fiecare dimineață în ultimele 10.000 de zile, probabil că va răsări și mâine dimineață”. Alteori, acestea pot fi mai nuanțate, cum ar fi „X% din familii au specii separate geografic, cu variante de culoare, deci există o șansă de Y% ca lebede negre nedescoperite să existe”.[11]

Programele de învățare automată pot îndeplini sarcini fără a fi programate în mod explicit pentru a face acest lucru. Implică ideea unor calculatoare care învață din datele furnizate astfel încât să îndeplinească anumite sarcini. Pentru sarcini simple atribuite computerelor, este posibil să se programeze algoritmi care spun mașinii cum să execute toți pașii necesari pentru a rezolva problema în cauză; din partea computerului, nu este nevoie de învățare. Pentru sarcini mai avansate, poate fi o provocare pentru un om să creeze manual algoritmii necesari. În practică, se poate dovedi mai eficient să ajute mașina să-și dezvolte propriul algoritm, mai degrabă decât ca programatorii umani să specifice fiecare pas necesar.[12]

Disciplina învățării automate folosește diverse abordări pentru a învăța computerele să îndeplinească sarcini în care nu este disponibil un algoritm pe deplin satisfăcător. În cazurile în care există un număr mare de răspunsuri potențiale, o abordare este de a eticheta unele dintre răspunsurile corecte ca fiind valide. Acestea pot fi apoi folosite ca date de antrenament pentru computer pentru a îmbunătăți algoritmul (algoritmii) pe care îl folosește pentru a determina răspunsurile corecte. De exemplu, pentru a instrui un sistem pentru sarcina de recunoaștere digitală a caracterelor, a fost adesea folosit setul de date MNIST de cifre scrise de mână.[12]

Abordări de învățare nesupravegheată: TextRank

Un algoritm de extragere a frazelor cheie este TextRank. (1) În timp ce metodele supravegheate au câteva proprietăți frumoase, cum ar fi capacitatea de a produce reguli interpretabile pentru caracteristicile care caracterizează o expresie cheie, ele necesită, de asemenea, o cantitate mare de date de antrenament. Sunt necesare multe documente cu expresii cheie cunoscute. În plus, antrenamentul pe un anumit domeniu tinde să personalizeze procesul de extracție la acel domeniu, astfel încât clasificatorul rezultat nu este neapărat portabil, așa cum demonstrează unele dintre rezultatele lui Turney. Extragerea nesupravegheată a frazelor cheie elimină nevoia de date de antrenament. Ea abordează problema dintr-un unghi diferit. În loc să încerce să învețe caracteristici explicite care caracterizează expresiile cheie, algoritmul TextRank exploatează structura textului în sine pentru a determina expresiile cheie care apar „centrale" pentru text, în același mod în care PageRank selectează paginile web importante. Amintiți-vă că acest lucru se bazează pe noțiunea de „prestigiu" sau „recomandare" din rețelele sociale. În acest fel, TextRank nu se bazează deloc pe date de antrenament anterioare, ci mai degrabă poate fi rulat pe orice bucată de text arbitrară și poate produce rezultate pur și simplu pe baza proprietăților intrinseci ale textului. Astfel algoritmul este ușor de portat în noi domenii și limbaje.

TextRank [13] este un algoritm de clasare bazat pe grafice de uz general pentru NLP. În esență, rulează PageRank pe un grafic special conceput pentru o anumită sarcină NLP. Pentru extragerea frazelor cheie, construiește un grafic folosind un set de unități de text ca noduri. Muchiile se bazează pe o anumită măsură a asemănării semantice sau lexicale între nodurile unității de text. Spre deosebire de PageRank, marginile sunt de obicei nedirecționate și pot fi ponderate pentru a reflecta un grad de similitudine. Odată construit graficul, acesta este folosit pentru a forma o matrice stocastică, combinată cu un factor de amortizare (ca în „modelul de surfer aleatoriu"), iar clasificarea peste noduri este obținută prin găsirea vectorului propriu corespunzător valorii proprii 1 (adică, distribuţia staţionară a mersului aleator pe grafic).

Nodurile ar trebui să corespundă cu ceea ce vrem să clasăm. Potențial, am putea face ceva similar cu metodele supravegheate și am crea un nod pentru fiecare unigramă, bigramă, trigramă etc. Cu toate acestea, pentru a menține graficul mic, autorii decid să clasifice unigramele individuale într-un prim pas, apoi să includă un al doilea pas care îmbină unigramele adiacente bine clasate pentru a forma expresii cu mai multe cuvinte. Acest lucru are un efect secundar frumos de a ne permite să producem fraze cheie de lungime arbitrară. De exemplu, dacă clasificăm

unigramele și constatăm că „procesare" „avansat", „limbaj" și „natural" obțin toate ranguri înalte, atunci ne-am uita la textul original și vom vedea că aceste cuvinte apar consecutiv și creează în final expresia cheie folosind toate patru împreună. Rețineți că unigramele plasate în grafic pot fi filtrate printr-o parte a vorbirii. Autorii au descoperit că adjectivele și substantivele sunt cele mai bune de inclus. Astfel, unele cunoștințe lingvistice intră în joc în acest pas.

Marginile sunt create pe baza apariției simultane a cuvintelor în această aplicație a TextRank. Două noduri sunt conectate printr-o muchie dacă unigramele apar într-o fereastră de dimensiune N în textul original. N este de obicei în jur de 2-10. Astfel, „natural" și „limbaj" ar putea fi legate într-un text despre NLP. „Natural" și „procesare" ar fi, de asemenea, conectate, deoarece ambele ar apărea în același șir de N cuvinte. Aceste margini se bazează pe noțiunea de „coeziune a textului" și pe ideea că cuvintele care apar unul lângă altul sunt probabil legate într-un mod semnificativ și se „recomandă" reciproc cititorului.

Deoarece această metodă clasifică pur și simplu nodurile individuale, avem nevoie de o modalitate de a limita sau de a produce un număr limitat de fraze cheie. Tehnica aleasă este să setați un număr T să fie o fracțiune specificată de utilizator din numărul total de noduri din grafic. Apoi, nodurile/unigramele de sus T sunt selectate pe baza probabilităților lor staționare. Se aplică apoi un pas de postprocesare pentru a îmbina instanțele adiacente ale acestor unigrame T. Ca rezultat, vor fi produse mai multe sau mai puține expresii cheie finale, dar numărul ar trebui să fie aproximativ proporțional cu lungimea textului original.

Inițial, nu este clar de ce aplicarea PageRank unui grafic de co-ocurență ar produce expresii cheie utile. O modalitate de a gândi la asta este următoarea. Un cuvânt care apare de mai multe ori într-un text poate avea mai mulți vecini concomitenți. De exemplu, într-un text despre învățarea automată, unigrama „învățare" poate apărea împreună cu „mașină", „supravegheată", „nesupravegheată" și „semi-supravegheată" în patru propoziții diferite. Astfel, nodul „învățare" ar fi un „hub" central care se conectează la aceste alte cuvinte modificatoare. Rularea PageRank/TextRank pe grafic are probabil ca „învățare" să se claseze foarte bine. În mod similar, dacă textul conține expresia „clasificare supravegheată", atunci ar exista o margine între „supravegheat" și „clasificare". Dacă „clasificarea" apare mai multe alte locuri și, prin urmare, are mulți vecini, importanța sa ar contribui la importanța „supravegheată". Dacă ajunge la un rang înalt, va fi selectat ca una dintre primele T unigrame, împreună cu „învățare" și probabil

„clasificare". În pasul final de post-procesare, vom ajunge apoi cu expresii cheie „învățare supravegheată" și „clasificare supravegheată".

Pe scurt, graficul de co-ocurență va conține regiuni dens conectate pentru termeni care apar des și în contexte diferite. O plimbare aleatorie pe acest grafic va avea o distribuție staționară care atribuie probabilități mari termenilor din centrele clusterelor. Acest lucru este similar cu paginile web dens conectate care sunt clasate foarte bine de PageRank. Această abordare a fost utilizată și în rezumarea documentelor, analizată mai jos.

TextRank și LexRank

Abordarea nesupravegheată a rezumatului este, de asemenea, destul de asemănătoare în spirit cu extragerea nesupravegheată a frazelor cheie și ocolește problema datelor costisitoare de antrenament. (1) Unele abordări de rezumare nesupravegheate se bazează pe găsirea unei propoziții „centroid", care este vectorul cuvântului mediu al tuturor propozițiilor din document. Apoi propozițiile pot fi clasificate în funcție de asemănarea lor cu această propoziție centroid.

O modalitate mai bazată pe principii de a estima importanța propoziției este utilizarea mersurilor aleatorii și a centralității vectorului propriu. LexRank [14] este un algoritm în esență identic cu TextRank și ambii folosesc această abordare pentru rezumarea documentelor. Cele două metode au fost dezvoltate de grupuri diferite în același timp, iar LexRank s-a concentrat pur și simplu pe rezumat, dar puteau fi la fel de ușor utilizate pentru extragerea expresiilor cheie sau orice altă sarcină de clasare NLP.

Atât în LexRank, cât și în TextRank, un grafic este construit prin crearea unui nod pentru fiecare propoziție din document.

Marginile dintre propoziții se bazează pe o formă de similitudine semantică sau de suprapunere a conținutului. În timp ce LexRank folosește asemănarea cosinus a vectorilor TF-IDF, TextRank folosește o măsură foarte similară, bazată pe numărul de cuvinte pe care două propoziții le au în comun (normalizate de lungimea propozițiilor). Lucrarea LexRank a explorat utilizarea marginilor neponderate după aplicarea unui prag valorilor cosinus, dar a experimentat și utilizarea marginilor cu ponderi egale cu scorul de similaritate. TextRank folosește scoruri de similaritate continue ca ponderi.

În ambii algoritmi, propozițiile sunt ordonate prin aplicarea PageRank la graficul rezultat. Un rezumat se formează prin combinarea propozițiilor de top, folosind un prag sau o lungime limită pentru a limita dimensiunea rezumatului.

Merită remarcat faptul că TextRank a fost aplicat rezumatului exact așa cum este descris aici, în timp ce LexRank a fost folosit ca parte a unui sistem de sumarizare mai mare (MEAD) care combină scorul LexRank (probabilitate staționară) cu alte caracteristici precum poziția și lungimea propoziției folosind o combinație liniară cu greutăți specificate de utilizator sau reglate automată. În acest caz, ar putea fi necesare unele documente de instruire, deși rezultatele TextRank arată că funcțiile suplimentare nu sunt absolut necesare.

O altă distincție importantă este că TextRank a fost folosit pentru rezumarea unui singur document, în timp ce LexRank a fost aplicat pentru rezumarea mai multor documente. Sarcina rămâne aceeași în ambele cazuri - doar numărul de propoziții din care să alegeți a crescut. Cu toate acestea, când rezumați mai multe documente, există un risc mai mare de a selecta propoziții duplicat sau extrem de redundante pentru a le plasa în același rezumat. Imaginați-vă că aveți un grup de articole de știri despre un anumit eveniment și doriți să realizați un rezumat. Este posibil ca fiecare articol să aibă multe propoziții similare și ați dori să includeți numai idei distincte în rezumat. Pentru a rezolva această problemă, LexRank aplică o etapă de post-procesare euristică, care creează un rezumat adăugând propoziții în ordinea clasamentului, dar renunță la orice propoziții care sunt prea asemănătoare cu cele deja plasate în rezumat. Metoda utilizată se numește Cross-Sentence Information Subsumption (CSIS).

Aceste metode funcționează pe baza ideii că propozițiile „recomandă" cititorului alte propoziții similare. Astfel, dacă o propoziție este foarte asemănătoare cu multe altele, va fi probabil o propoziție de mare importanță. Importanța acestei propoziții provine și din importanța propozițiilor care o „recomandă". Astfel, pentru a obține o poziție superioară și plasată într-un rezumat, o propoziție trebuie să fie similară cu multe propoziții care sunt, la rândul lor, similare cu multe alte propoziții. Acest lucru are sens intuitiv și permite aplicarea algoritmilor oricărui text nou arbitrar. Metodele sunt independente de domeniu și ușor de portat. Ne-am putea imagina caracteristicile care indică propoziții importante în domeniul științelor putând varia considerabil față de domeniul biomedical. Cu toate acestea, abordarea nesupravegheată bazată pe „recomandări" se aplică oricărui domeniu.

Referințe

(1) Bentley, Drew (2023). *Business intelligence și analitica în afaceri*, Traducere și adaptare de Nicolae Sfetcu. Editura MultiMedia Publishing, ISBN 978-606-033-776-8, DOI: 10.58679/MM70651

(2) Învățarea automată în inteligența artificială, în *Telework*, 27.04.2023, https://www.telework.ro/ro/invatarea-automata-in-inteligenta-artificiala/

Bibliografie

[1] Mitchell, Tom *(1997)*. *Machine Learning*. New York: McGraw Hill. *ISBN 0-07-042807-7*. *OCLC 36417892*. *Archived* from the original on 2020-04-07.

[2] George, Gerard; Haas, Martine R.; Pentland, Alex (2014). *"FROM THE EDITORS: BIG DATA AND MANAGEMENT"*. *The Academy of Management Journal*. **57** (2): 321–326. *ISSN 0001-4273*.

[3] Definiția „fără a fi programat în mod explicit” este adesea atribuită lui Arthur Samuel, care a inventat termenul de „învățare automată” în 1959, dar expresia nu se găsește literal în această publicație și poate fi o parafrază care a apărut mai târziu. Confer „Parafrazând Arthur Samuel (1959), întrebarea este: Cum pot calculatoarele să învețe să rezolve probleme fără a fi programate în mod explicit?” în Koza, John R.; Bennett, Forrest H.; Andre, David; Keane, Martin A. (1996). *"Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming". Artificial Intelligence in Design '96. Artificial Intelligence in Design '96.* Springer, Dordrecht. pp. 151–170. *doi:10.1007/978-94-009-0279-4_9*. *ISBN 978-94-010-6610-5*.

[4] Hu, Junyan; Niu, Hanlin; Carrasco, Joaquin; Lennox, Barry; Arvin, Farshad (2020). *"Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning"*. IEEE Transactions on Vehicular Technology. **69** (12): 14413–14423. *doi:10.1109/tvt.2020.3034800*. *ISSN 0018-9545*. *S2CID 228989788*.

[5] Yoosefzadeh-Najafabadi, Mohsen; Hugh, Earl; Tulpan, Dan; Sulik, John; Eskandari, Milad (2021). *"Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean?"*. *Front. Plant Sci.* **11**: 624273. doi:10.3389/fpls.2020.624273. PMC 7835636. PMID 33510761.

[6] Bishop, C. M. *(2006), Pattern Recognition and Machine Learning, Springer,* ISBN 978-0-387-31073-2

[7] Machine learning and pattern recognition "can be viewed as two facets of the same field."[6]: vii

[8] Friedman, Jerome H. *(1998)*. *"Data Mining and Statistics: What's the connection?"*. *Computing Science and Statistics*. **29** (1): 3–9.

[9] *"What is Machine Learning?"*. *www.ibm.com*. *Archived* from the original on 2021-08-13

[10] Zhou, Victor (2019-12-20). *"Machine Learning for Beginners: An Introduction to Neural Networks"*. *Medium*. *Archived* from the original on 2022-03-09

[11] Domingos, Pedro (September 22, 2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books. ISBN 978-0465065707, Chapter 6, Chapter 7.

[12] *Ethem Alpaydin (2020). Introduction to Machine Learning (Fourth ed.). MIT*. pp. xix, 1–3, 13–18. *ISBN 978-0262043793*.

[13] Rada Mihalcea and Paul Tarau, 2004: *TextRank: Bringing Order into Texts*, Department of Computer Science University of North Texas *"Archived copy" (PDF).*

[14] Güneş Erkan and Dragomir R. Radev: *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization* [1]